Basic QTL mapping

Karl Broman

Biostatistics and Medical Informatics
University of Wisconsin – Madison
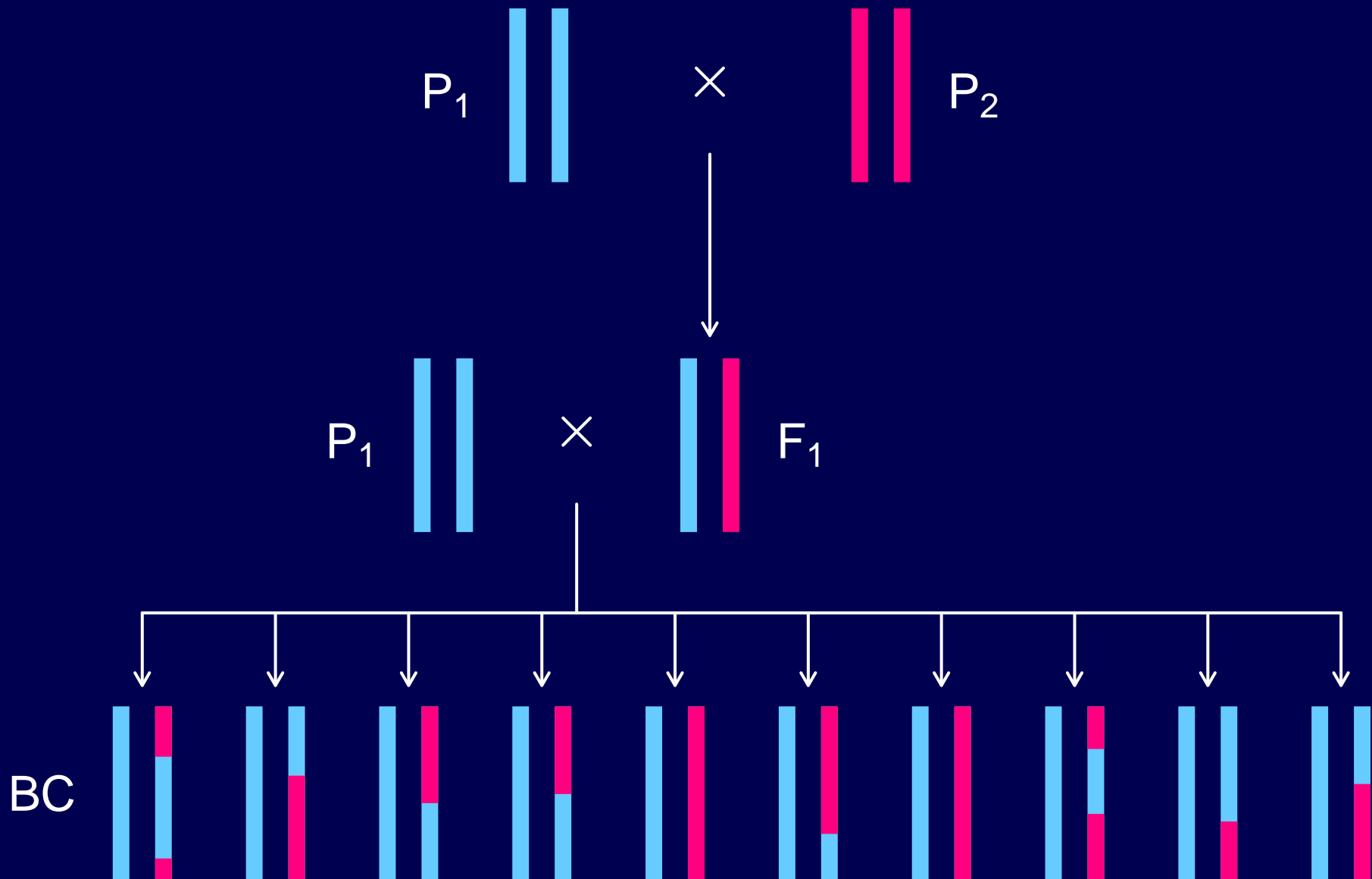
rqtl.org

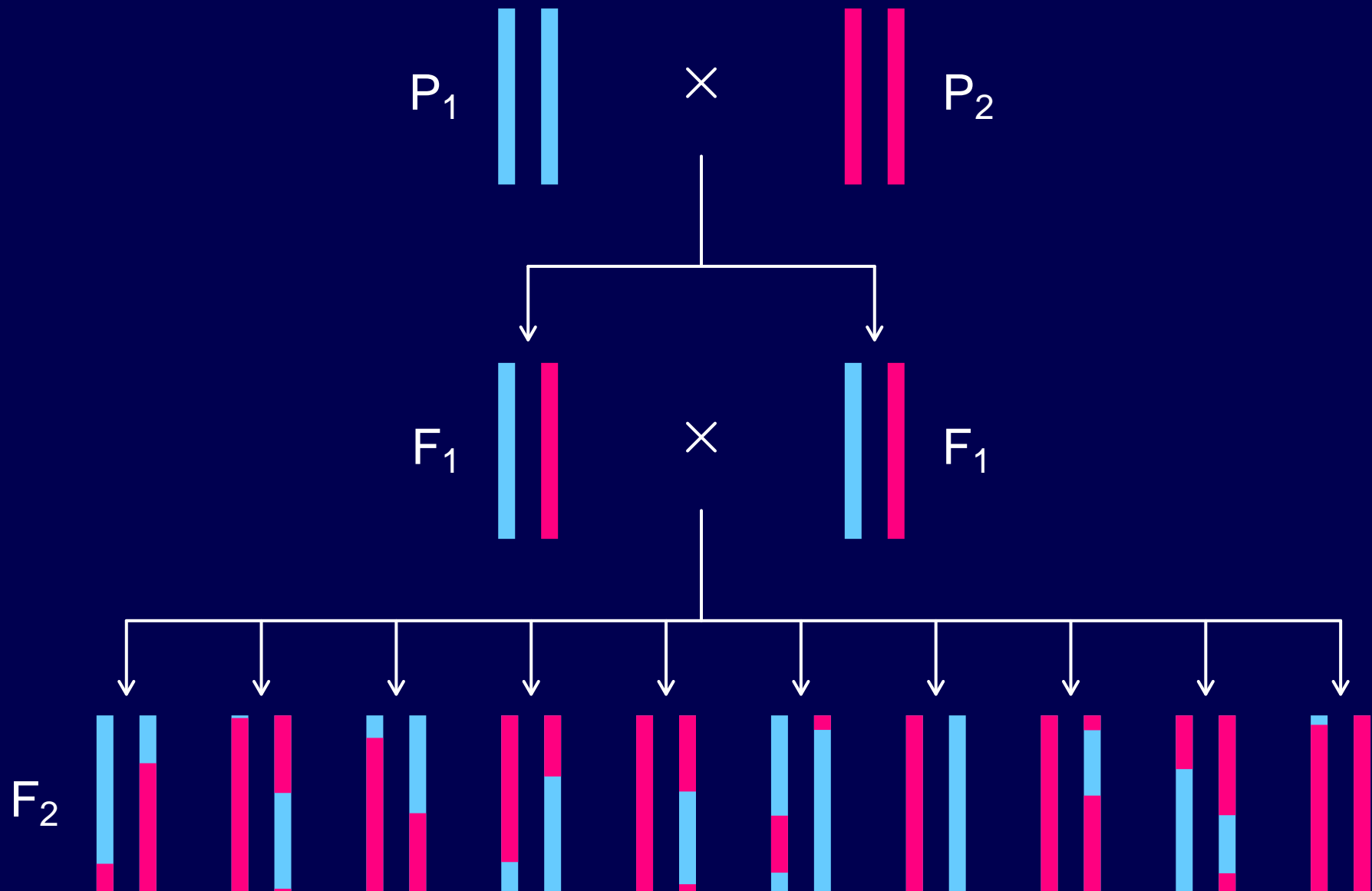kbroman.org

github.com/kbroman

@kwbroman

Backcross

Intercross

Goals

- Identify quantitative trait loci (QTL)
  (and interactions among QTL)

- Interval estimates of QTL location

- Estimated QTL effects

→ R

- R, RStudio, and R/qtl

- `read.cross()`

- `summary(), plot()`

- `nind(), nmar(), totmar(), nchr(), nphe()`

ANOVA at marker loci

- Also known as marker regression.

- Split mice into groups according to genotype at a marker.

- Do a t-test / ANOVA.

- Repeat for each marker.
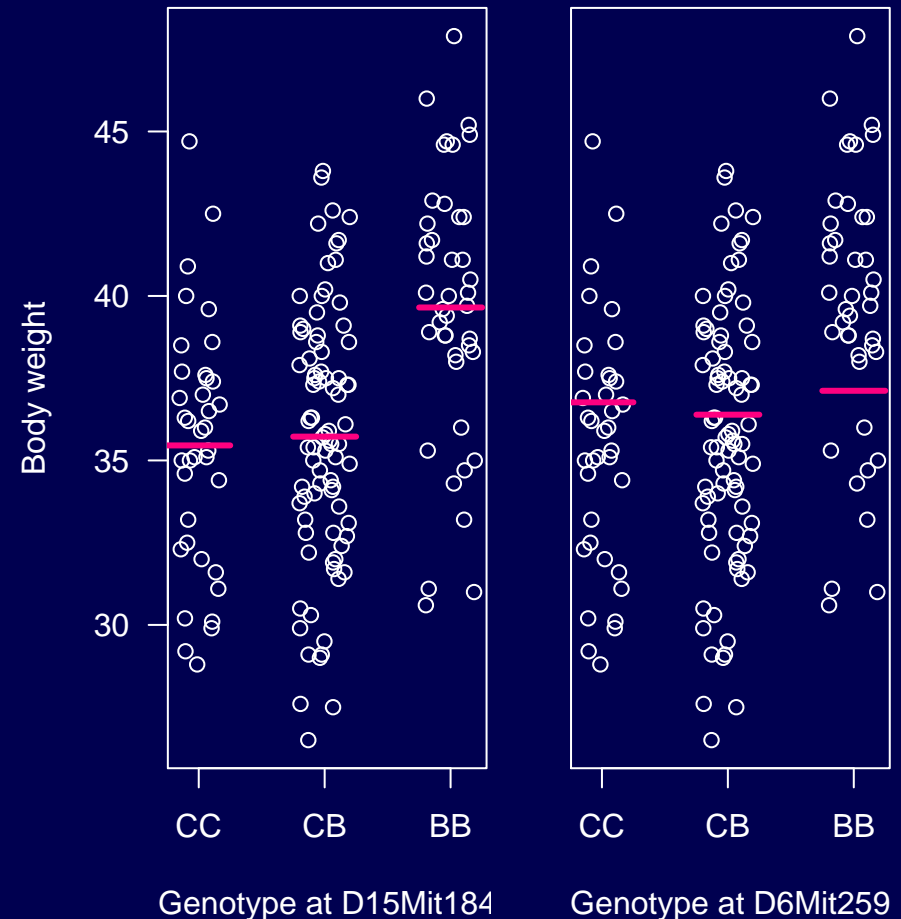
ANOVA at marker loci

Advantages

- Simple.

- Easily incorporates covariates.

- Easily extended to more complex models.

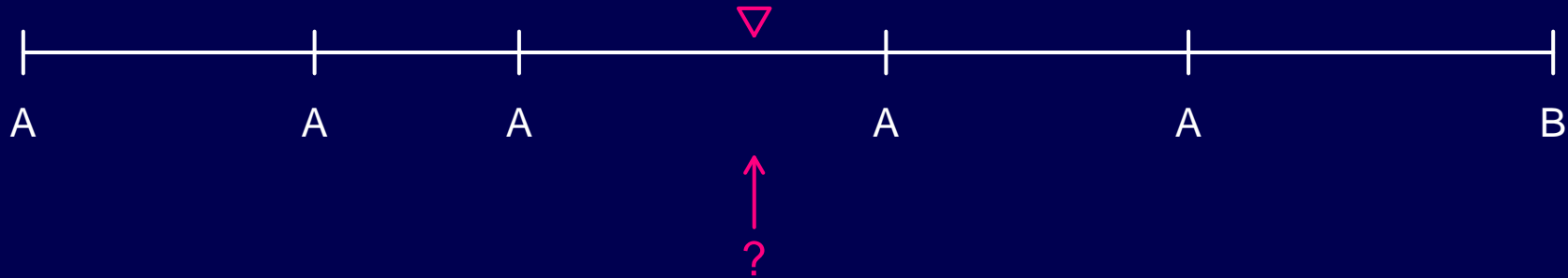- Doesn't require a genetic map.

Disadvantages

- Must exclude individuals with missing genotype data.

- Imperfect information about QTL location.

- Suffers in low density scans.

- Only considers one QTL at a time.

Interval mapping

Lander & Botstein (1989)

- Assume a single QTL model.

- Each position in the genome, one at a time, is posited as the putative QTL.

- Let q = the unobserved QTL genotype
  Assume $y|q \sim N(\mu_q, \sigma)$

- We don't know q, but we can calculate $\Pr(q \mid \text{marker data})$

- Estimate μ_q, σ by *maximum likelihood* using an iterative EM algorithm
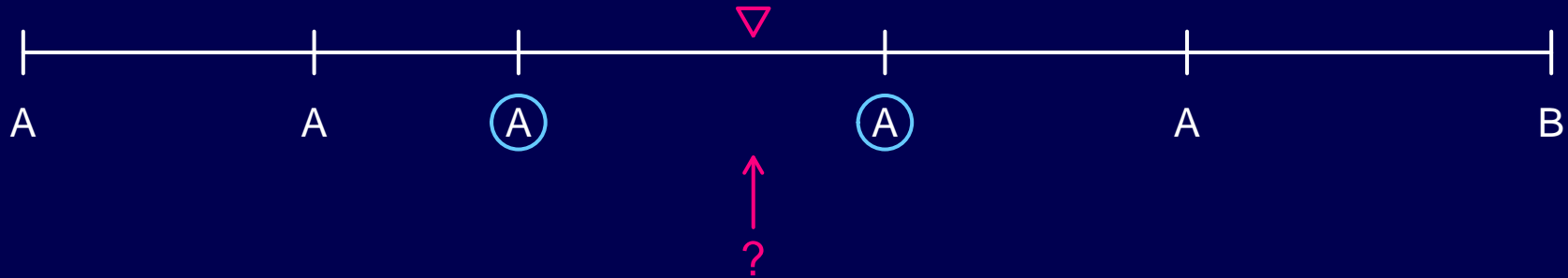
Genotype probabilities

Calculate $\Pr(q \mid \text{marker data})$, assuming

- No crossover interference

- No genotyping errors

Or use the hidden Markov model (HMM) technology

- To allow for genotyping errors

- To incorporate dominant markers

- (Still assume no crossover interference.)

Genotype probabilities

Calculate $\Pr(\text{q} \mid \text{marker data})$, assuming

- No crossover interference

- No genotyping errors

Or use the hidden Markov model (HMM) technology

- To allow for genotyping errors

- To incorporate dominant markers

- (Still assume no crossover interference.)
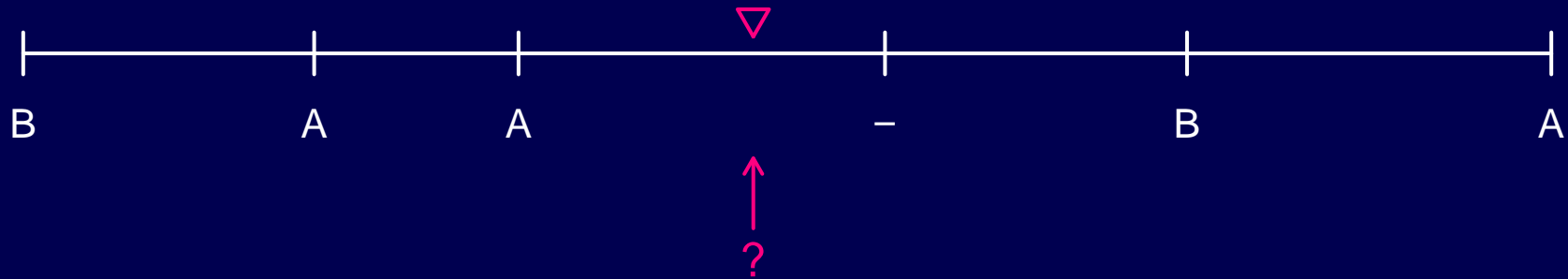
Genotype probabilities

Calculate $\Pr(q \mid \text{marker data})$, assuming

- No crossover interference
- No genotyping errors

Or use the hidden Markov model (HMM) technology

- To allow for genotyping errors
- To incorporate dominant markers
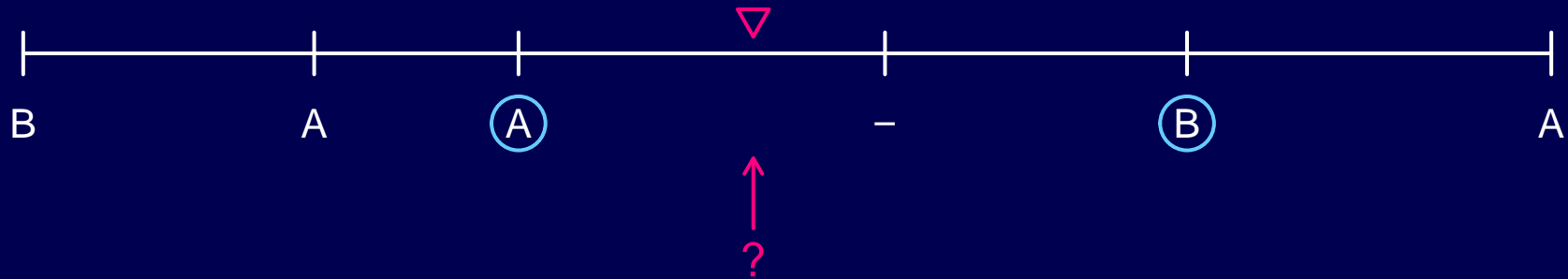- (Still assume no crossover interference.)

Genotype probabilities

Calculate $\Pr(q \mid \text{marker data})$, assuming

- No crossover interference
- No genotyping errors

Or use the hidden Markov model (HMM) technology

- To allow for genotyping errors
- To incorporate dominant markers
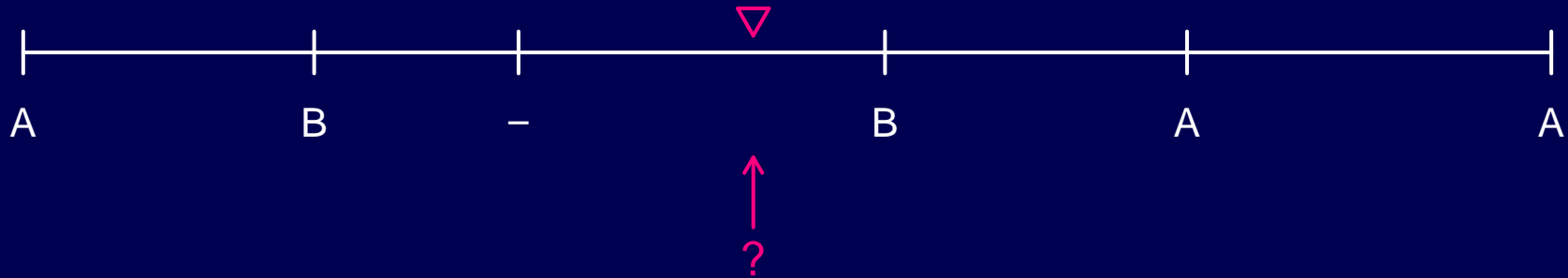- (Still assume no crossover interference.)
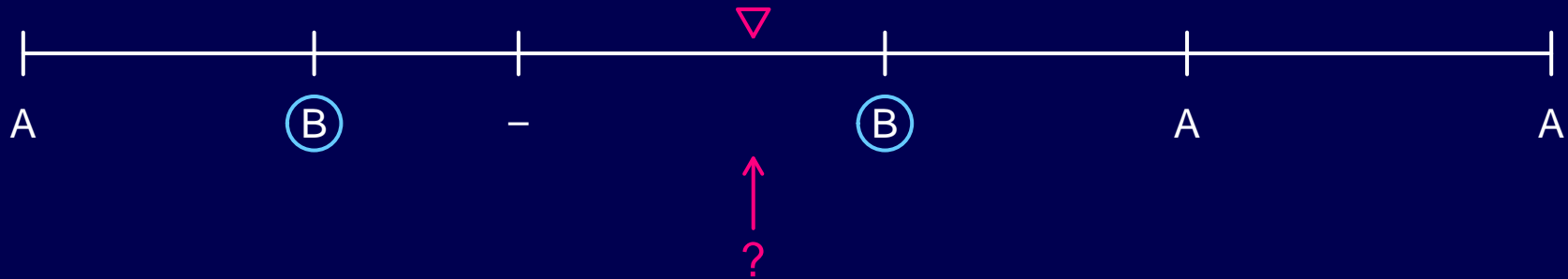
Genotype probabilities

Calculate $\Pr(q \mid \text{marker data})$, assuming

- No crossover interference

- No genotyping errors

Or use the hidden Markov model (HMM) technology

- To allow for genotyping errors

- To incorporate dominant markers

- (Still assume no crossover interference.)

Calculate $\Pr(q \mid \text{marker data})$, assuming

- No crossover interference
- No genotyping errors

Or use the hidden Markov model (HMM) technology

- To allow for genotyping errors
- To incorporate dominant markers
- (Still assume no crossover interference.)

LOD scores

The LOD score is a measure of the strength of evidence for the presence of a QTL at a particular location.

$\text{LOD}(\lambda) = \log_{10}$ likelihood ratio comparing the hypothesis of a QTL at position λ versus that of no QTL

$$= \log_{10} \left\{ \frac{\Pr(y | \text{QTL at } \lambda, \hat{\mu}_{0\lambda}, \hat{\mu}_{1\lambda}, \hat{\sigma}_\lambda)}{\Pr(y | \text{no QTL}, \hat{\mu}, \hat{\sigma})} \right\}$$

$\hat{\mu}_{0\lambda}, \hat{\mu}_{1\lambda}, \hat{\sigma}_\lambda$ are the MLEs, assuming a single QTL at position λ.

No QTL model: The phenotypes are independent and identically distributed (iid) $N(\mu, \sigma^2)$.

→ R

- `calc.genoprob()`

- `scanone()`

- `iplotScanone()` from R/qtlcharts

Interval mapping

Advantages

- Takes proper account of missing data.

- Allows examination of positions between markers.

- Gives improved estimates of QTL effects.

- Provides pretty graphs.

Disadvantages

- Increased computation time.

- Requires specialized software.

- Difficult to generalize.

- Only considers one QTL at a time.

LOD thresholds

Large LOD scores indicate evidence for the presence of a QTL

Question: How large is large?

LOD threshold = 95 %ile of distr'n of max LOD, genome-wide, if there are
no QTLs anywhere

Derivation:
- Analytical calculations (L & B 1989)
- Simulations (L & B 1989)
- Permutation tests (Churchill & Doerge 1994)
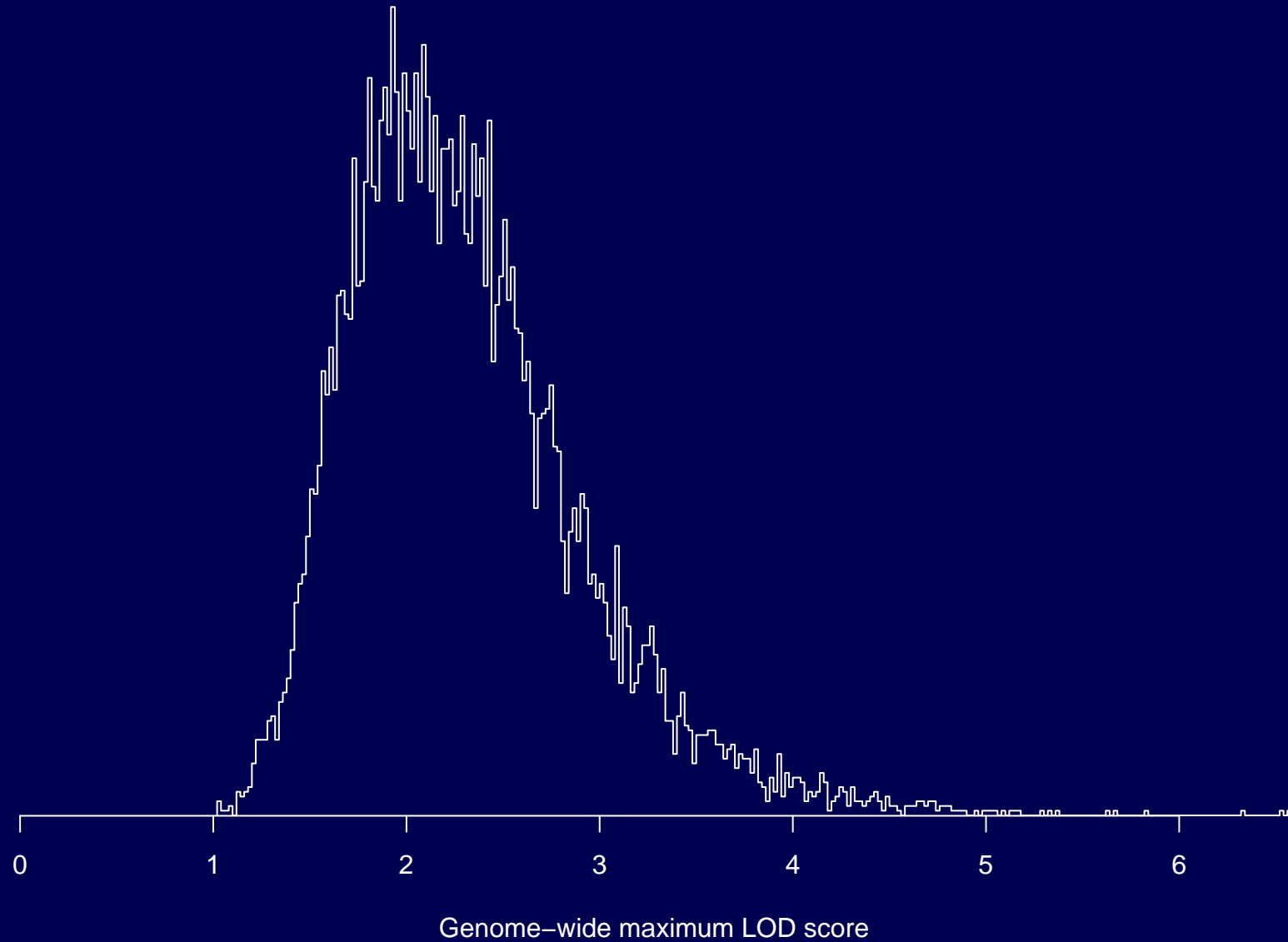
Null distribution of the LOD score

- Null distribution derived by computer simulation of backcross with genome of typical size.

- Dashed curve: distribution of LOD score at any one point.

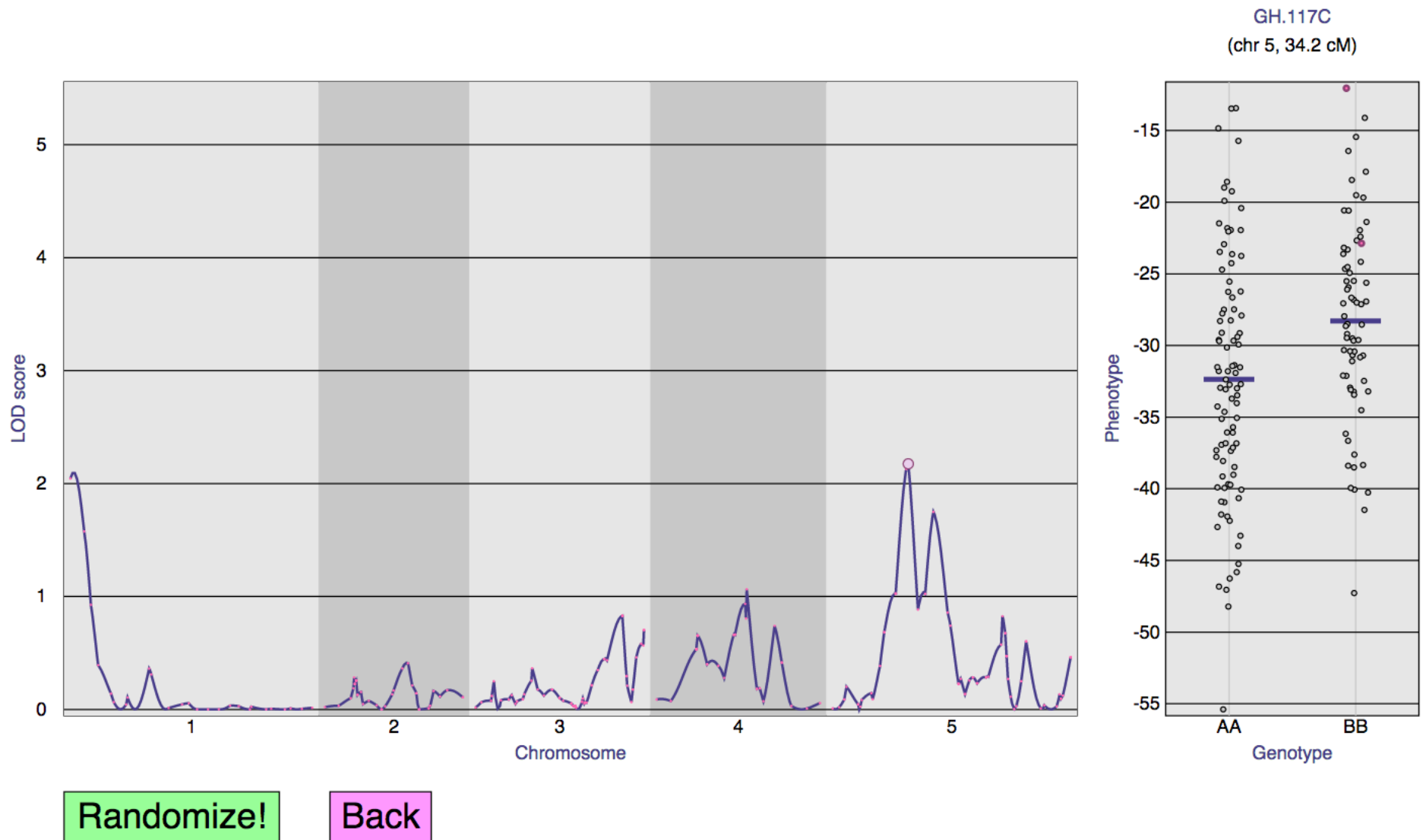- Solid curve: distribution of maximum LOD score, genome-wide.
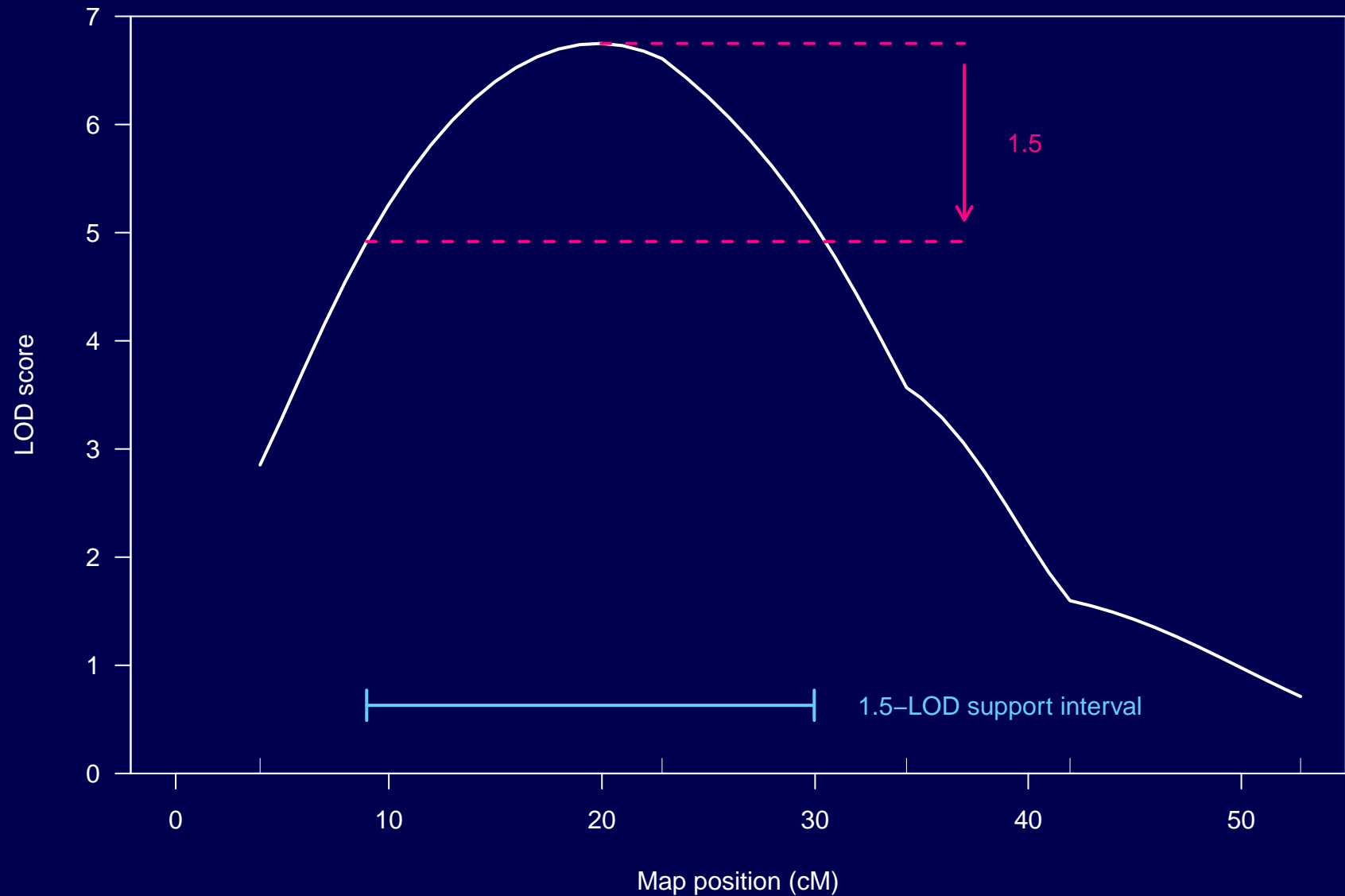

LOD score
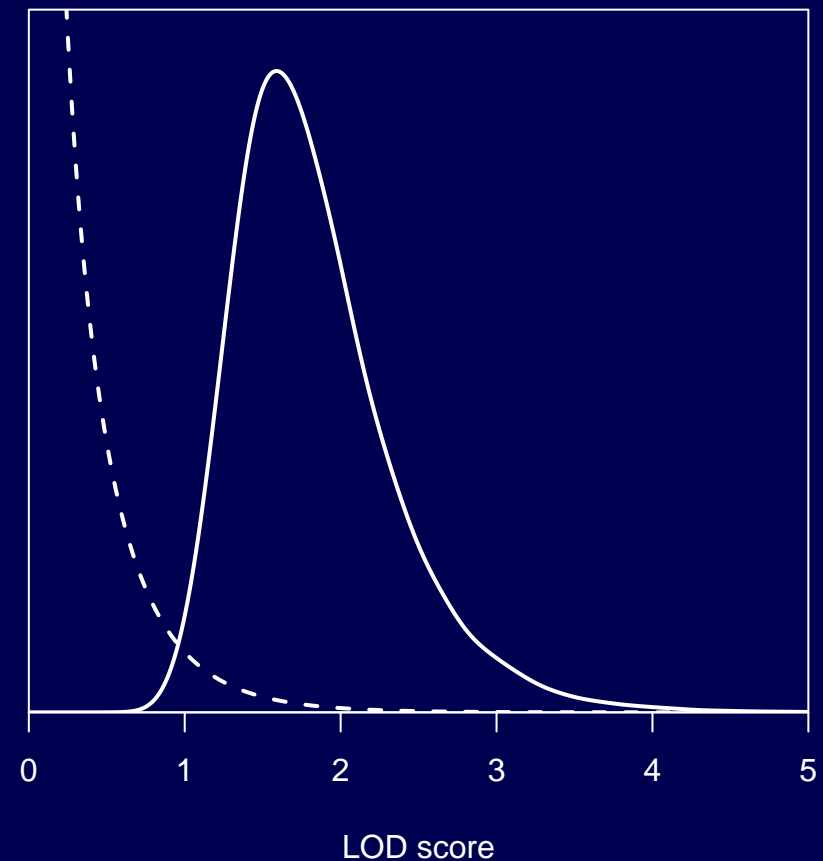
Permutation test
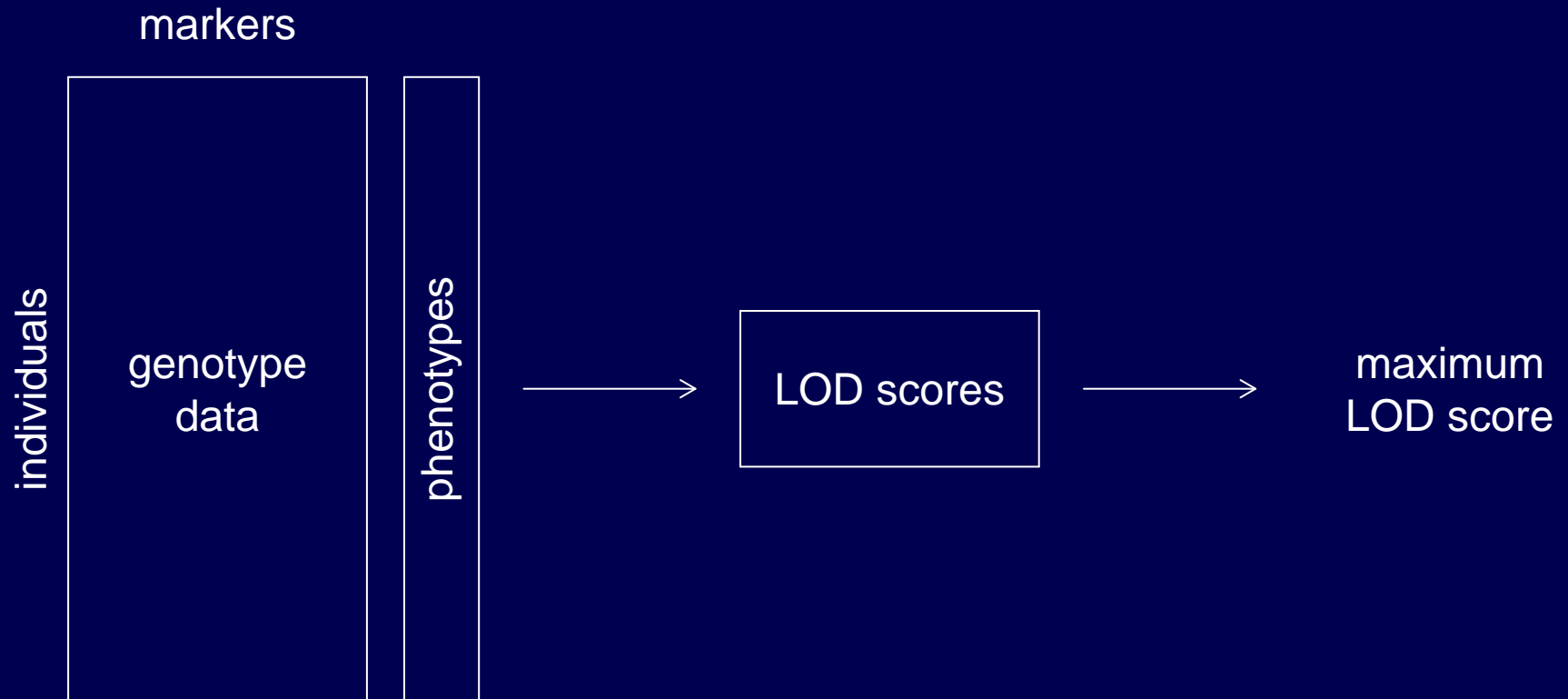
Permutation results

Genome−wide maximum LOD score

Interactive plot

$\rightarrow$ R

- `scanone()` for permutations

LOD support intervals

→ R

- lodint()

- bayesint()

Haley-Knott regression

A quick approximation to Interval Mapping.

$$E(y_i|q_i) = \mu_q$$

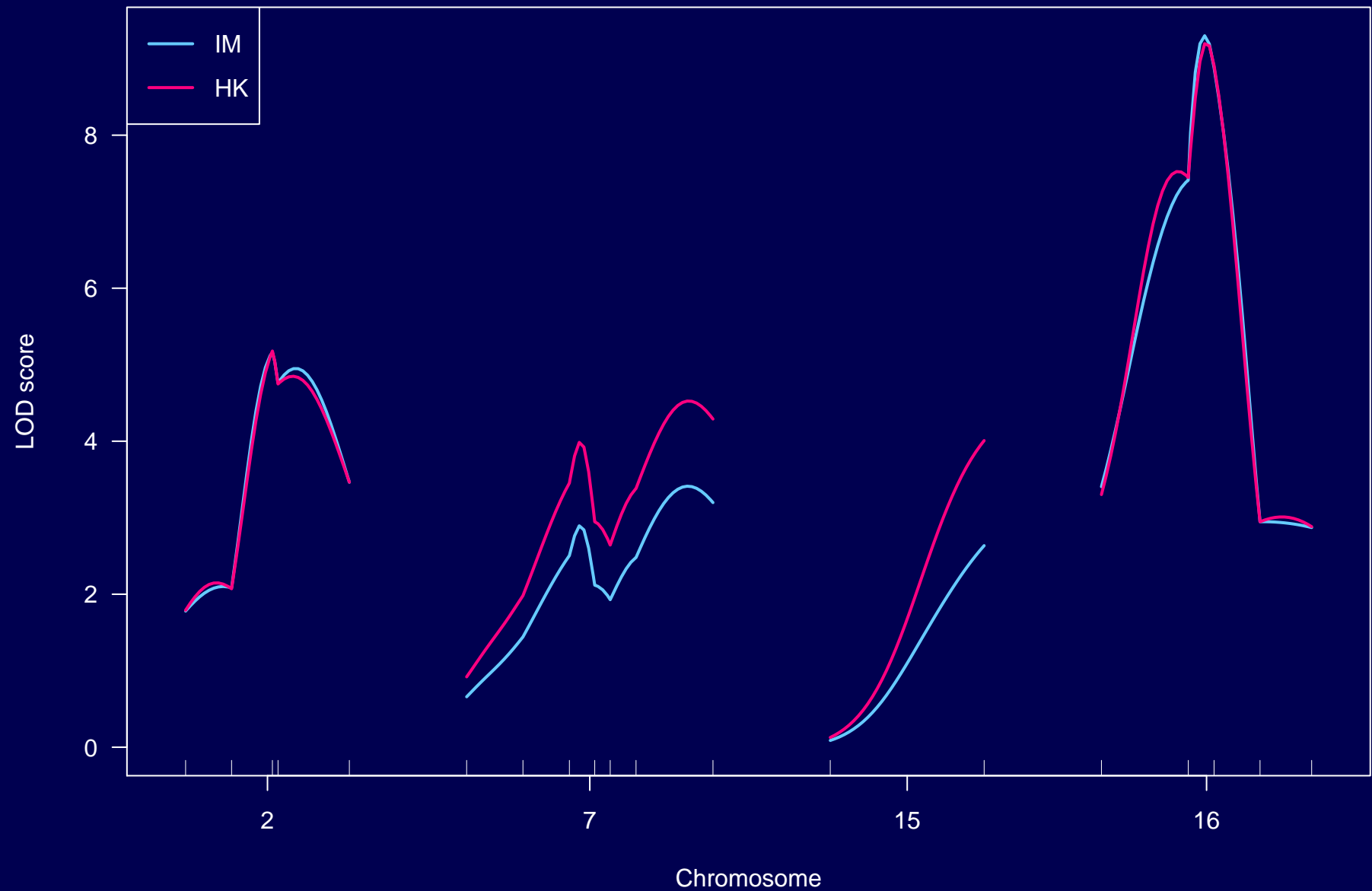$$E(y_i|M_i) = E[\, E(y_i|q_i)\, |M_i] = \sum_j \Pr(q = j|M_i)\mu_j$$

$$= \sum_j p_{ij}\mu_j$$

Regress y on p_i, pretending the residual variation is normally distributed (with constant variance).
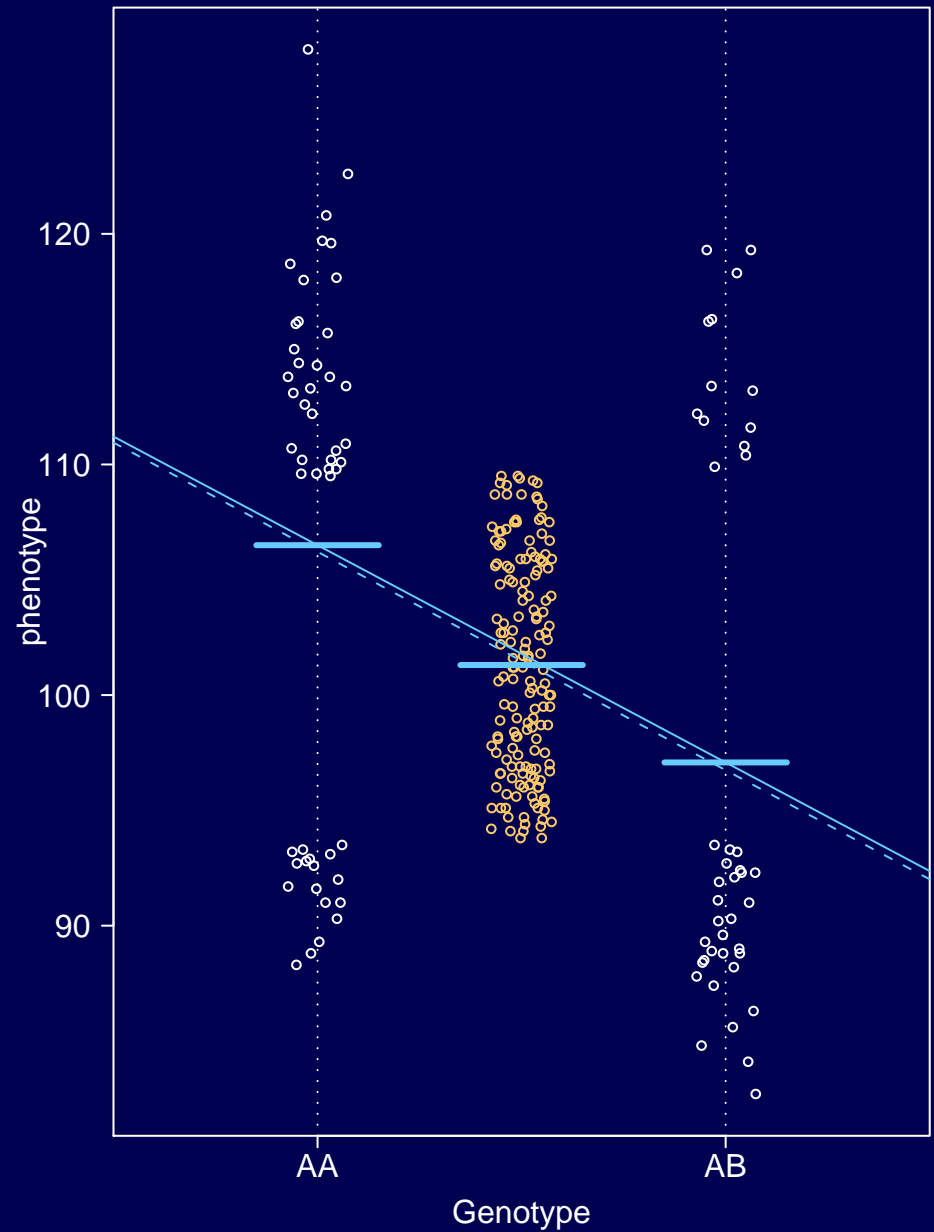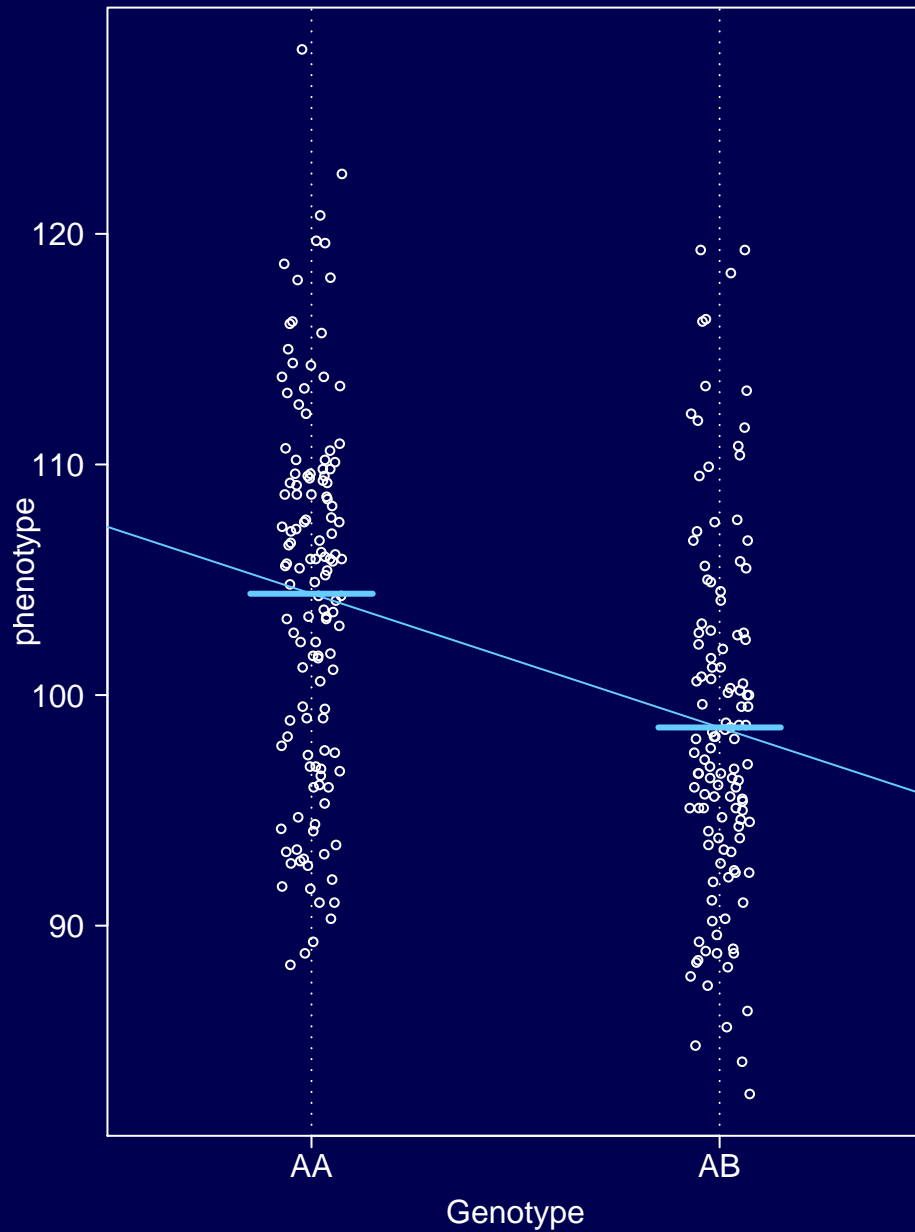
$$LOD = \frac{n}{2}\log_{10}\left(\frac{RSS_0}{RSS_1}\right)$$
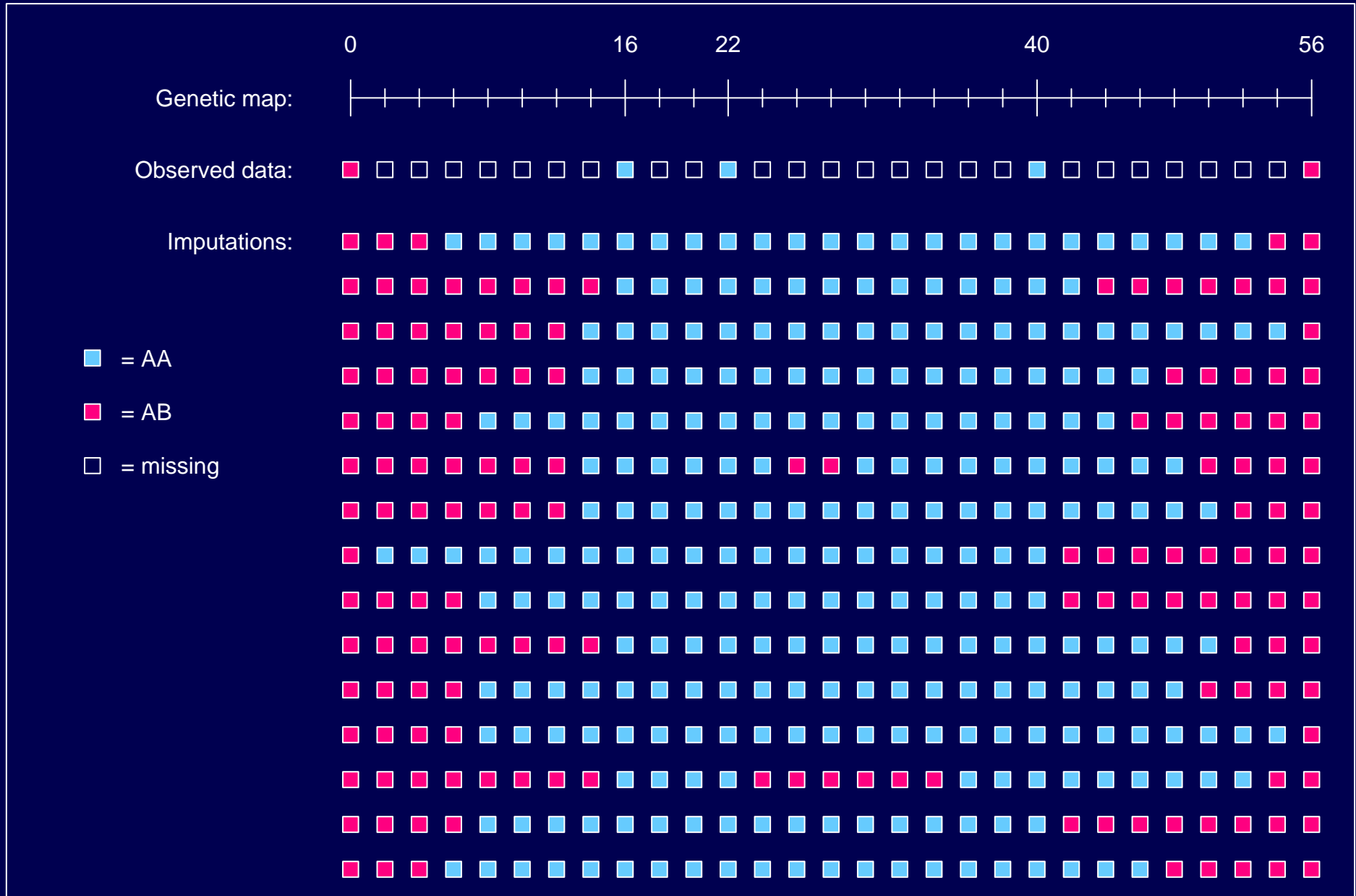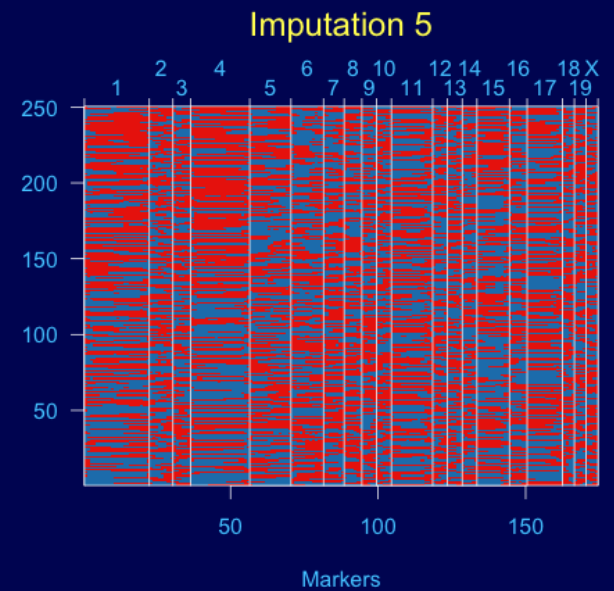
→ R

- scanone() with method="hk"

Haley-Knott results

H-K with selective genotyping

Multiple imputation

Multiple imputations
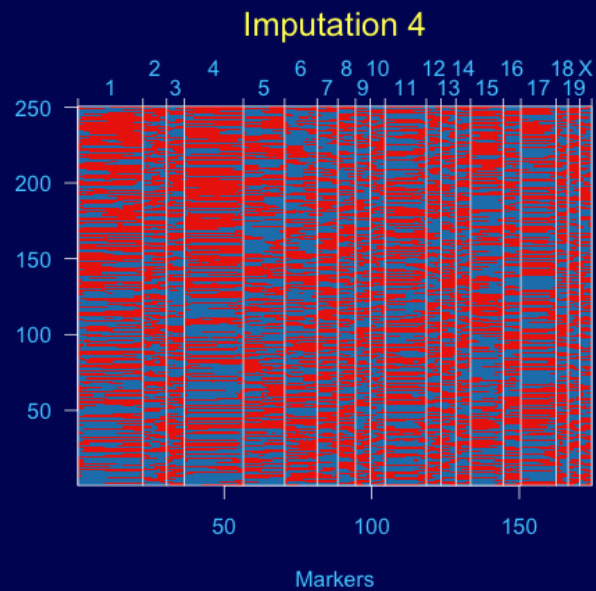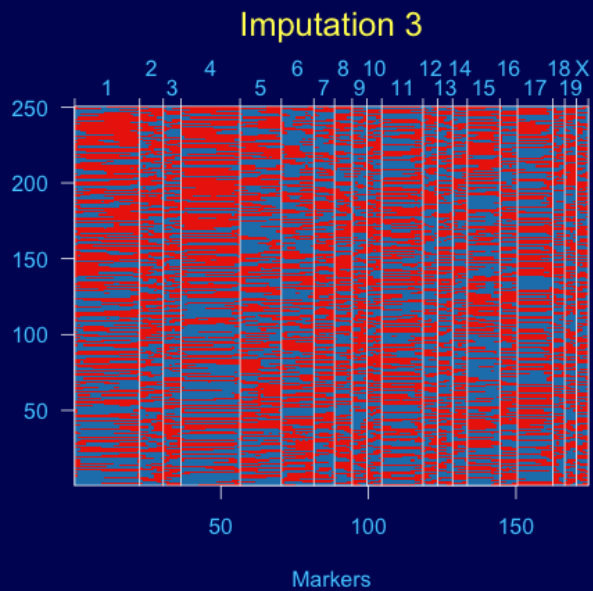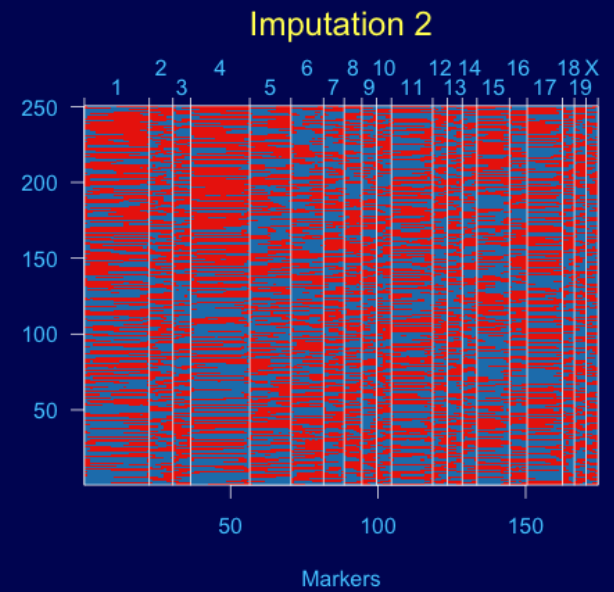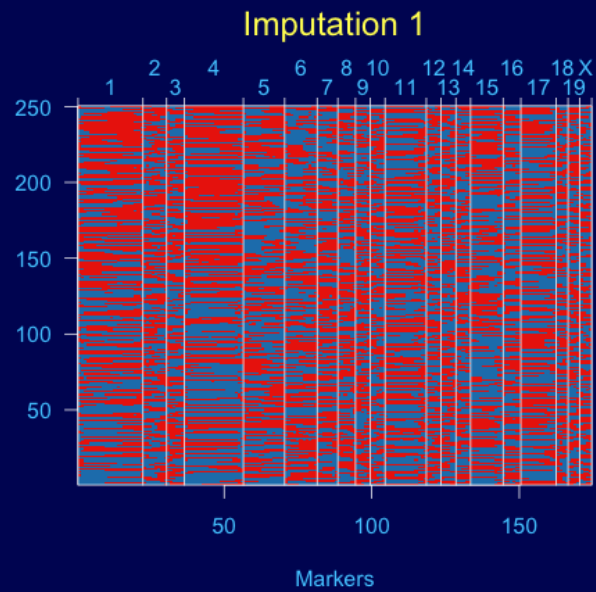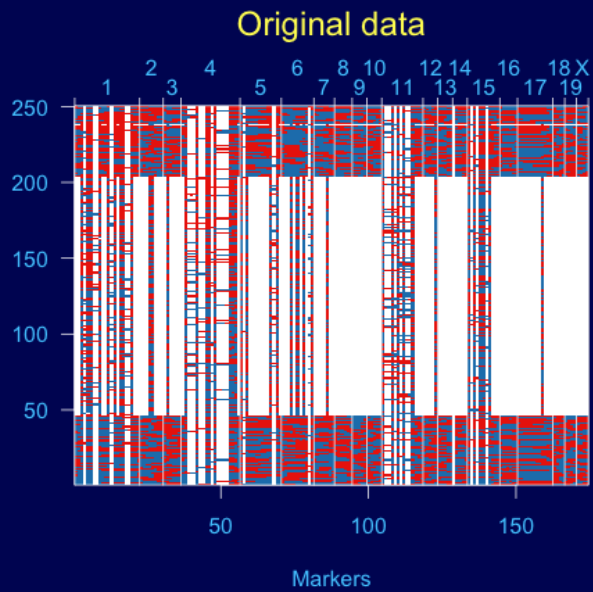
Imputation LOD curves
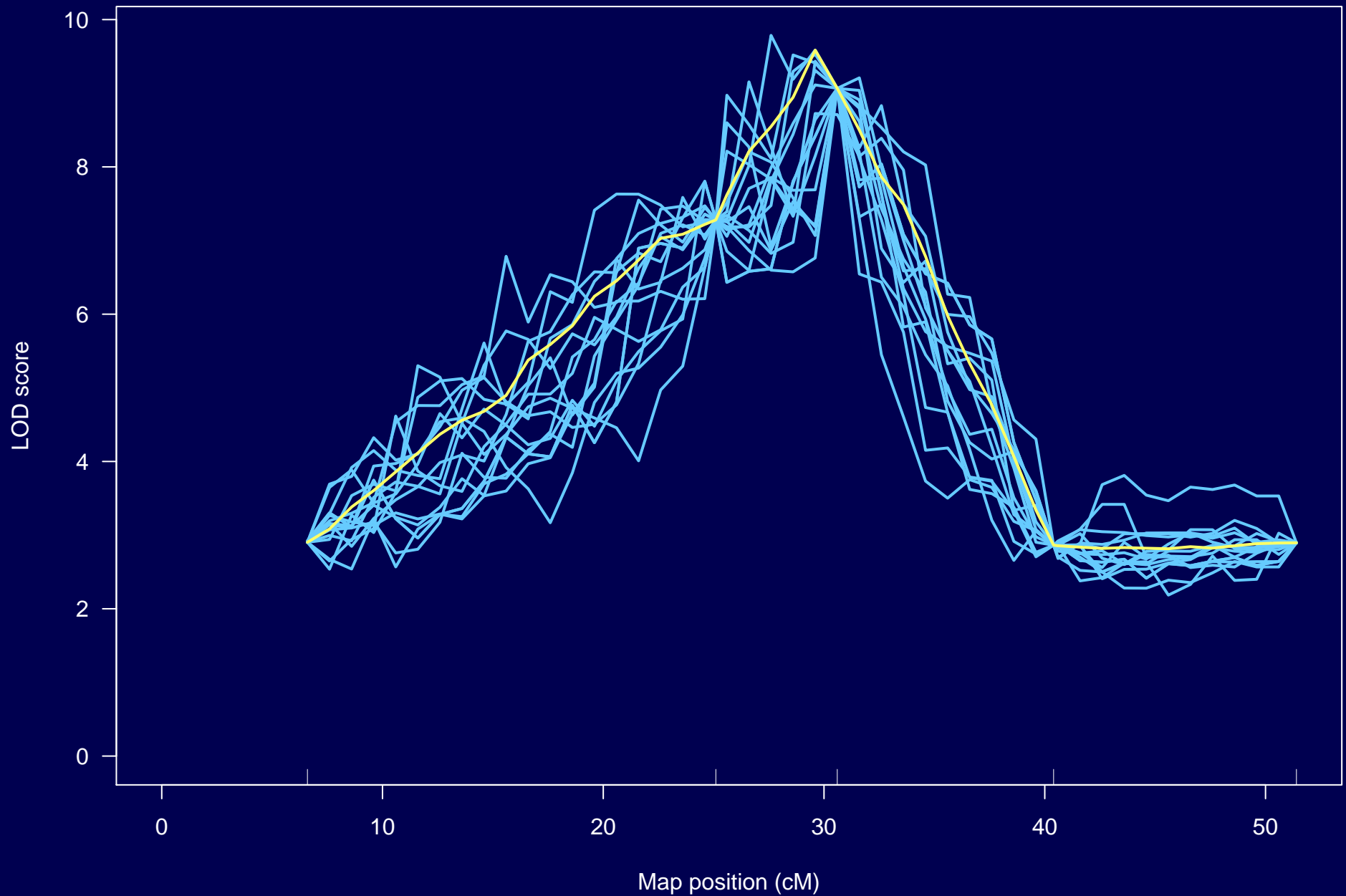
→ R

- `sim.geno()`

- `scanone()` with `method="imp"`

Summary comparison

Approach	Speed	Extensibility	Stability	Missing data	Parallelization
HK	++	+	+	−	++
EM	+	−	−	+	−
Imputation	−	+	+	+	+

Non-normal traits

- Standard interval mapping assumes normally distributed residual variation. (Thus the phenotype distribution is a mixture of normals.)

- In reality: we see dichotomous traits, counts, skewed distributions, outliers, and all sorts of odd things.

- Interval mapping, with LOD thresholds derived from permutation tests, generally performs just fine anyway.

- Alternatives to consider:

  – Nonparametric approaches (Kruglyak & Lander 1995)
  – Transformations (*e.g.*, log, square root, normal quantiles)
  – Specially-tailored models (*e.g.*, a generalized linear model, the Cox proportional hazard model, and the two-part model in Broman 2003)

→ R

- `nqrank()`

- `scanone()` with `model="binary"` or `model="np"`

Covariates

- Examples: treatment, sex, age, weight

- Control residual variation $\rightarrow$ increase power

- Look for QTL $\times$ covariate interactions

Additive covariate

$$H_0 : y = \mu + \beta_x x + \epsilon$$
$$H_a : y = \mu + \beta_x x + \beta_q q + \epsilon$$

- If covariate has strong effect on the phenotype, accounting for it can give improved power to detect QTL.

- In permutations, keep phenotype and covariate together

- Use care when the covariate is another phenotype

Additive covariate
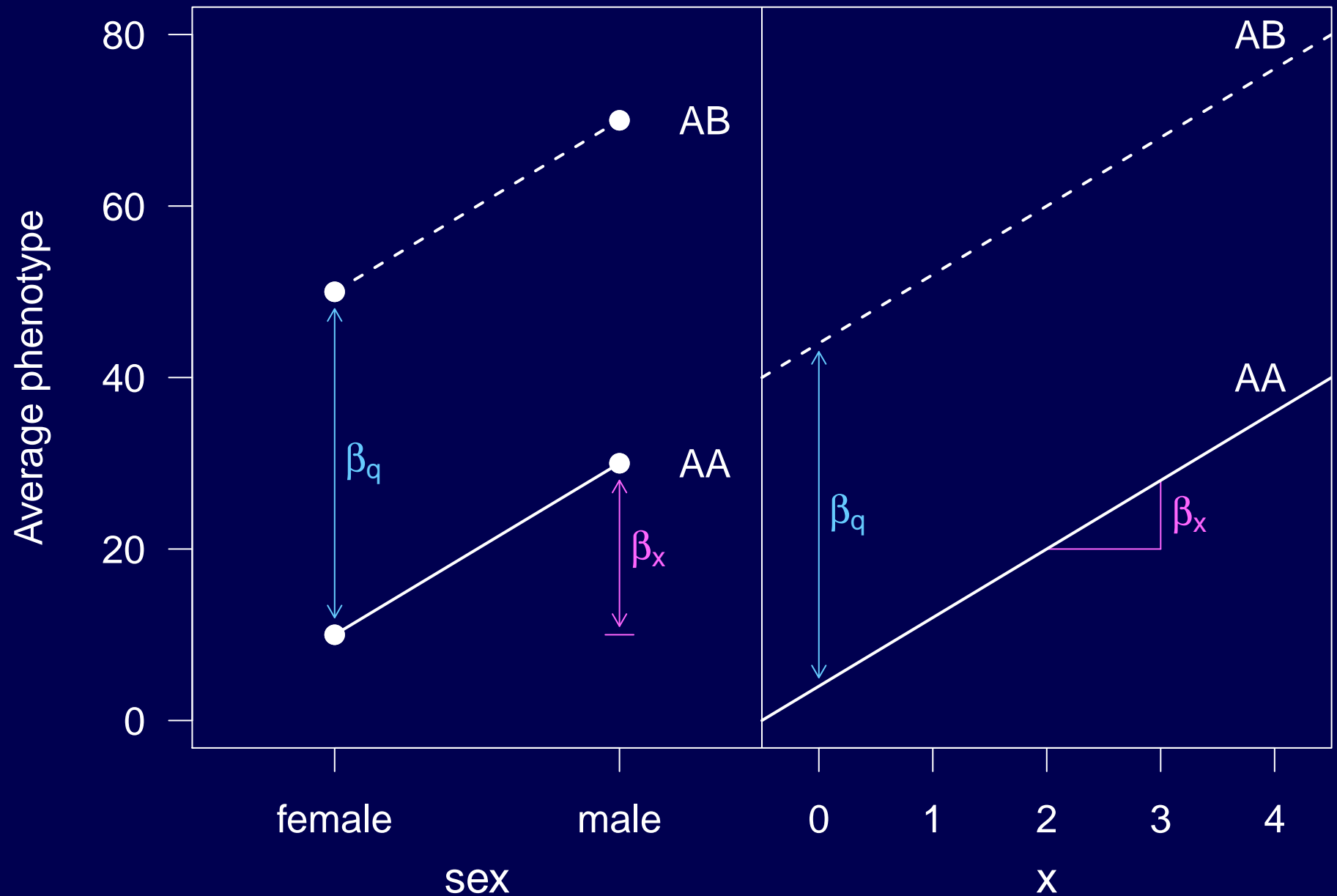
Adjust then scan?

- Consider adjusted phenotype $y' = y/x$

- The QTL model is $(y/x) = \mu + \beta_q q + \epsilon$

- Equivalently

$$y = \left\{ \begin{array}{ll} \mu\, x + \epsilon' & \text{if } q = 0 \\ (\mu + \beta_q)x + \epsilon' & \text{if } q = 1 \end{array} \right.$$
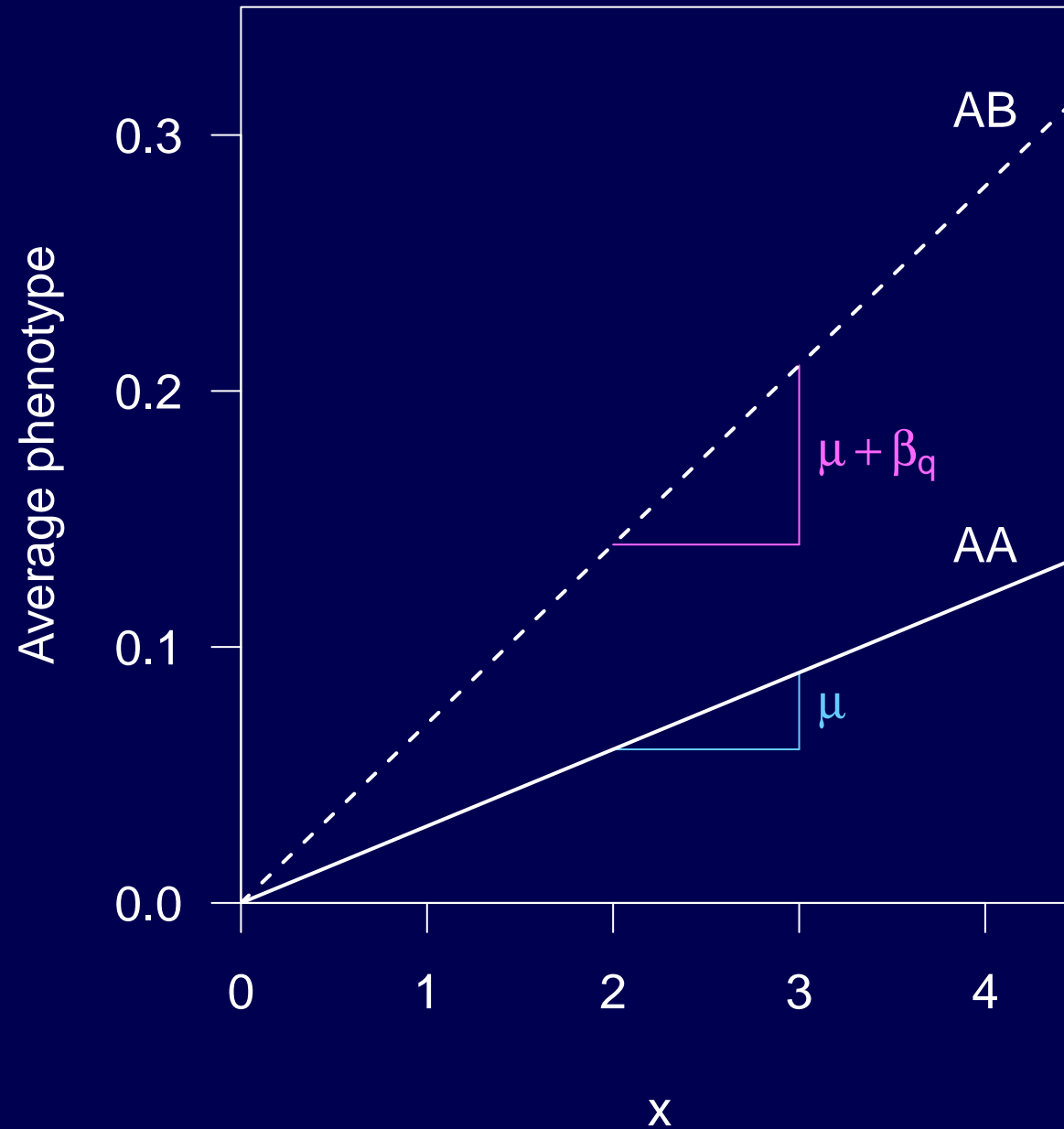
Interactive covariate

$$H_0 : y = \mu + \beta_x x + \epsilon$$
$$H_a : y = \mu + \beta_x x + \beta_q q + \epsilon$$
$$H_i : y = \mu + \beta_x x + \beta_q q + \gamma x q + \epsilon$$

Can consider 3 LOD scores:

- LOD_a comparing H_a and H_0
- LOD_f comparing H_i and H_0
- LOD_i comparing H_i and H_a

Interactive covariate
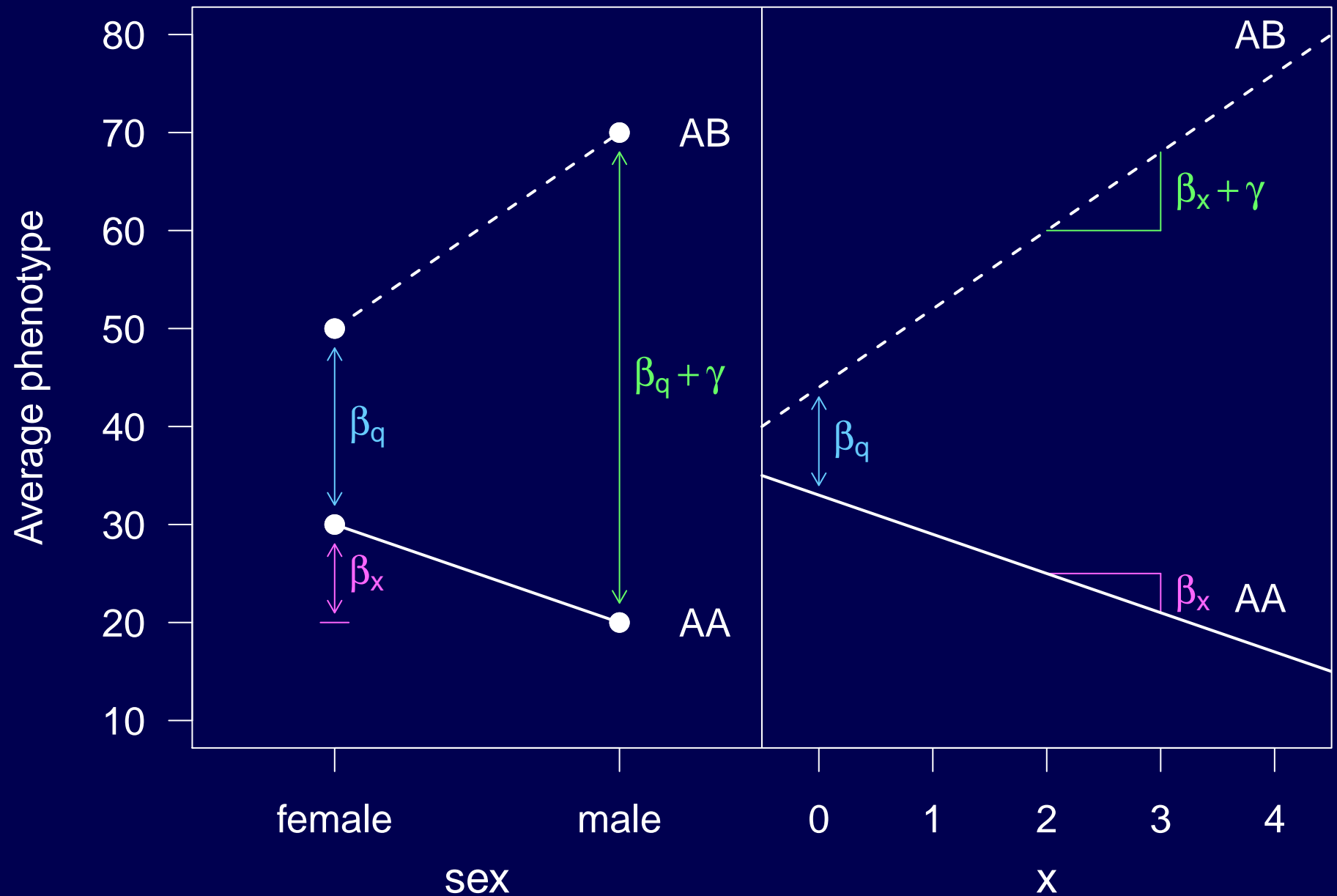
→ R

- `scanone()` with `addcovar` and `intcovar`

- `set.seed()` to do permutations

Split on sex?

- Informative, understandable

- But tempting to falsely conclude "sex-specific QTL"

- Absence of evidence is not *evidence of absence*.

- Use explicit test of QTL × sex interaction

Chromosome 6
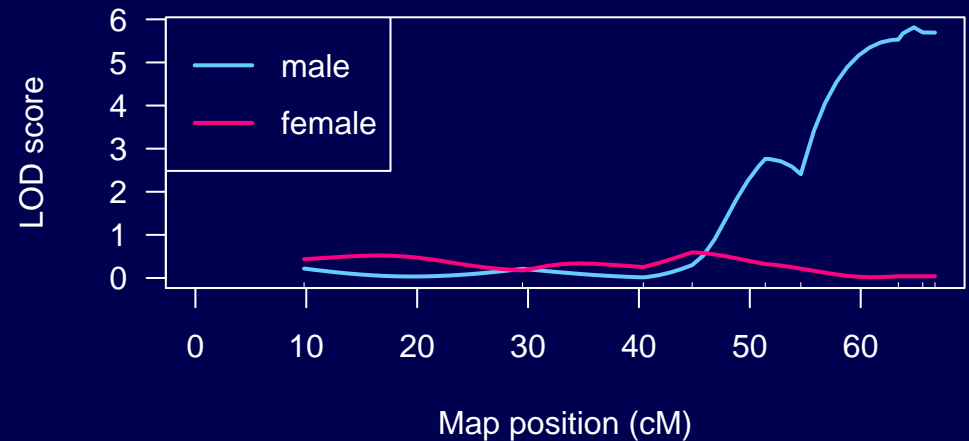
D6Mit373
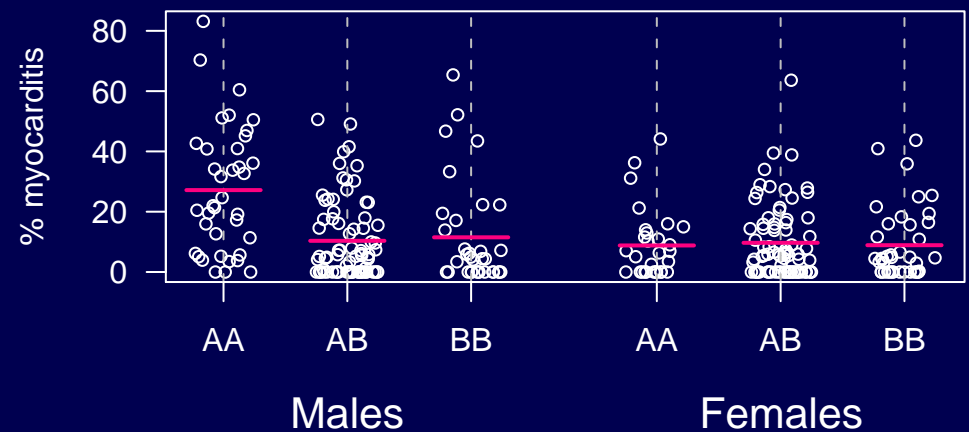
→ R

- `subset()` to split on sex

Data diagnostics

- Plot phenotypes

- Look for sample duplicates

- Look for excessive missing data

- Investigate segregation distortion

- Verify genetic maps/marker positions

- Look for genotyping errors

- Look at counts of crossovers

See Ch 3 in the R/qtl book, `rqtl.org/book`

References

- Broman KW (2001) Review of statistical methods for QTL mapping in experimental crosses. Lab Animal 30:44–52

  A review for non-statisticians.

- Lynch M, Walsh B (1998) *Genetics and analysis of quantitative traits*. Sinauer Associates, Sunderland, MA, chapter 15

  Chapter on QTL mapping.

- Lander ES, Botstein D (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185–199

  The seminal paper.

- Churchill GA, Doerge RW (1994) Empirical threshold values for quantitative trait mapping. Genetics 138:963–971

  LOD thresholds by permutation tests.

References

- Beavis WD (1994). The power and deceit of QTL experiments: Lessons from comparative QTL studies. In DB Wilkinson, (ed) 49th Ann Corn Sorghum Res Conf, pp 252–268. Amer Seed Trade Asso, Washington, DC.

  Discusses selection bias in estimated QTL effects.

- Broman KW (2003) Mapping quantitative trait loci in the case of a spike in the phenotype distribution. Genetics 163:1169–1175

  Two-part model; also discusses binary traits and non-parametric QTL mapping.

- Haley CS, Knott SA (1992) A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity 69: 315–324

  Haley-Knott regression

- Sen S, Churchill GA (2001) A statistical framework for quantitative trait mapping. Genetics 159: 371–387

  Multiple imputation

- Solberg LC, et al. (2004) Sex- and line-specific lineage inheritance of depression-like behavior in the rat. Mamm Genome 15:648–662

  Additive and interactive covariates.

- Broman KW et al (2006) The X chromosome in quantitative trait locus mapping. Genetics 174:2151–2158

References

- Broman KW, Speed TP (2002) A model selection approach for the identification of quantitative trait loci in experimental crosses. J Roy Stat Soc B 64:641–656

  Multiple-QTL model selection with additive QTL.

- Manichaikul A, Moon JY, Sen Ś, Yandell BS, Broman KW (2009) A model selection approach for the identification of quantitative trait loci in experimental crosses, allowing epistasis. Genetics 181:1077–1086

  Also account for epistasis.